New Perspectives in Astrophysical Cosmology

Second edition

Martin Rees
King's College, University of Cambridge

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, uk   http://www.cup.cam.ac.uk
40 West 20th Street, New York, NY 10011-4211, USA   http://www.cup.org
10 Stamford Road, Oakleigh, Melbourne 3166, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain

First published 1995
Second edition 2000

Printed in the United Kngdom at the University Press, Cambridge

Typeset in Hollander 10/15pt [VN]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data

Rees, Martin J., 1942–
Perspectives in astrophysical cosmology / Martin Rees. – 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0 521 64238 8 (hc.)
1. Cosmology. 2. Galaxies. 3. Nuclear astrophysics. I. Title.
QB981.R37 2000
523.1–dc21 99–21389 CIP

ISBN 0 521 64238 8 hardback

Contents

1
The cosmological framework

Introduction

Gravity, almost undetectable between laboratory-scale bodies, is the dominant force in astronomy and cosmology. The basic structures in our cosmic environment – stars, galaxies, and clusters of galaxies – all involve a balance between gravitational attraction and the disruptive effect of pressure or kinetic energy. Our entire observable universe may display a similar balance: the Hubble expansion is being slowed (and may perhaps eventually be braked to a halt) by the gravitational effect of its entire mass-energy.

The best-understood cosmic structures are the smaller ones: the individual stars. Stellar structures and life-cycles can be predicted theoretically, and tested empirically by observing large populations of stars, of differing ages, in the Milky Way. The Milky Way, the disc galaxy to which the Sun belongs, can be envisaged as a kind of ecological system in which stars are continually being born and dying, their gaseous content being recycled and chemically enriched as the evolution proceeds.

Our own Galaxy is typical of the galaxies distributed through the universe, which are the most conspicuous features of the cosmic scene. Why should the universe be full of these remarkable aggregates of stars and gas, typically $\sim 10^5$ light-years across and containing around 10^{11} stars? We do not yet have compelling physical reasons for the characteristic properties of galaxies, as we do for stars.

One reason why galaxies are harder to understand than stars is that their formation impinges on *cosmology*. Individual stars form, evolve, and die more or less regardless of what the universe does – initial cosmic conditions have left no traces on the complex gas dynamics that goes on within each galaxy. But that is not true for galaxies, which may have emerged, at an epoch when the entire universe was denser and perhaps very different, from inhomogeneities that were imprinted on the universe in its earliest phases.

Large-scale structure: how homogeneous is the universe?

In the perspective of the cosmologist, even entire galaxies are little more than 'points of light' which indicate how the material content of the universe is distributed, and how it moves. Galaxies are clustered: some in small groups (like our own Local Group, of which the Milky Way and the Andromeda galaxy are the dominant members), others in big clusters with hundreds of members. Moreover the clusters themselves are grouped in filamentary or sheet-like superclusters. In recent years, there has been great progress in quantifying the distribution of galaxies over the sky, and also in mapping out the three-dimensional structure. The latter task has entailed determining redshifts and distances for thousands of galaxies.

Figure 1 shows the major groupings of galaxies within our

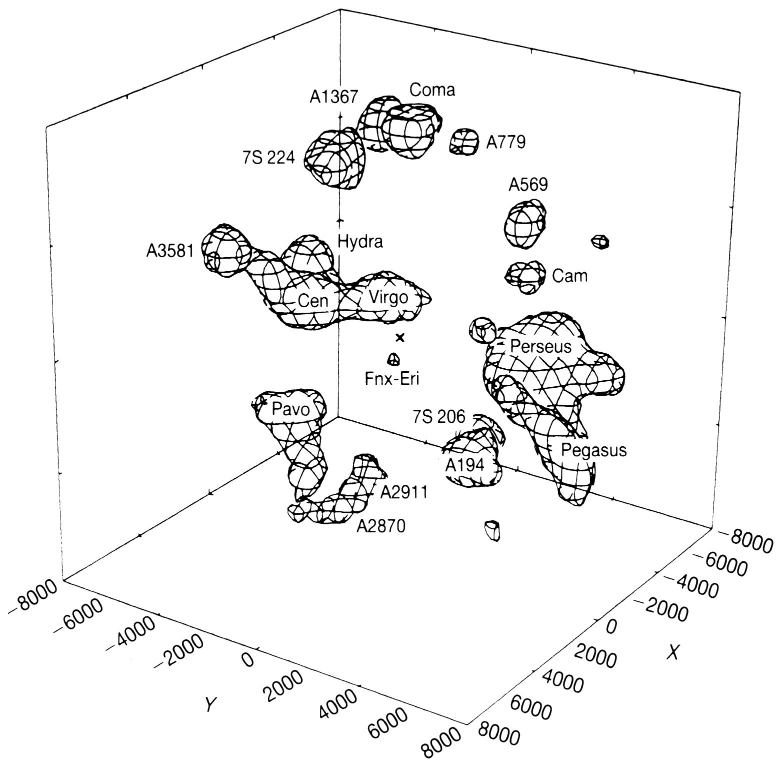

Figure 1

The most conspicuous clusters and superclusters within a cube, of dimensions around 3×10^8 light-years (10^8 pc), centred on our own Galaxy. There are also, of course, many galaxies more uniformly distributed in the space between these clusters. The linear dimensions of the region depicted here are about 2 per cent of the size of the part of the universe accessible to optical observations. This cube is probably large enough to provide a 'fair sample' of the contents of the universe: on larger scales the amplitude of inhomogeneities is much less than unity. (From Hudson, M. J. 1993, *Mon. Not. Roy. Astron. Soc.* **265**, 43 (Fig. 10).)

local part of the universe, out to a distance of around 3×10^8 light-years. At still greater distances, galaxies are distributed more uniformly over the sky. There is no evidence that big density contrasts extend to larger scales. A region of the size shown in this figure is therefore probably large enough to provide a fair sample of the contents of the universe.

Our universe certainly does not have a simple fractal structure, with clusters of clusters of clusters *ad infinitum*. There is a definite upper limit to the scale on which large-amplitude density inhomogeneities are observed. The *largest* structures with $(\delta\rho/\rho) \gtrsim 1$ are about 1 per cent of the Hubble radius: the typical nonlinear scale is around 0.3 per cent. The typical metric fluctuations due to clusters and superclusters – defined in dimensionless form as the gravitational energy per unit mass arising from the associated density enhancement, in units of c^2 – *have an amplitude of the order of* $Q = 10^{-5}$. The velocities relative to the Hubble flow induced by these structures are typically below $Q^{1/2}c = 1000$ km s^{-1}. The mass-equivalent of the kinetic energy associated with these so-called 'peculiar motions' is therefore only 10^{-5} of the rest mass. This number $Q = 10^{-5}$, a measure of the metric fluctuations in our universe, has an importance which will come up again later. Its smallness implies that the present local dynamics of cosmic inhomogeneities such as clusters and superclusters can be validly approximated by Newtonian gravity. Even more importantly, it justifies the relevance of the simple theoretical models for a homogeneous isotropic universe. These models date back to the 1920s.

The first relativistic models for a homogeneous expanding universe were found by Friedmann[1] before Hubble[2] discovered the recession of the nebulae. Hubble's work, which showed that the universe did not resemble Einstein's earlier static model, stimulated further studies of relativistic cosmology by Lemaitre, Tolman, and others. But the data were then – and remained for several decades – too sparse to indicate whether any of these idealised models fitted the real universe, still less to discriminate among them.

High-redshift objects

Hubble's work suggested that the galaxies would have been crowded together in the past, and emerged from some kind of 'beginning'. But he had no direct evidence for cosmic evolution: indeed the steady-state theory,[3] proposed in 1948 as a tenable alternative to the 'big bang', envisaged continuous creation of new matter and new galaxies, so that despite the expansion the overall cosmic scene never changed.

To discern any cosmic evolutionary trend, one must probe objects so far away that their light set out when the universe was significantly younger. This entails studying objects billions of light-years away with substantial redshifts. A programme to measure the cosmic deceleration was pursued from the 1950s onwards with the 200-inch Palomar telescope.[4] But the results were inconclusive, partly because normal galaxies are not luminous enough to be detectable at sufficiently large redshifts. It was Ryle and his colleagues from radio astronomy,[5] in the late 1950s, who found the first real evidence that the universe was indeed evolving. Radio telescopes could pick up emission from some unusual 'active' galaxies (which are now believed to be harbouring massive black holes in their centres) even when they were too far away to be seen with optical telescopes. One cannot determine the redshift or distance of such sources from radio measurements alone, but Ryle assumed that, statistically at least, the ones appearing faint were more distant than those appearing intense. He counted the numbers with various apparent intensities, and found that there were too many apparently faint ones – in other words, sources at large distances – compared with the number of brighter and closer ones. This was discomforting to the 'steady statesmen', but compatible with an evolving universe if galaxies were more prone to undergo violent outbursts in the remote past, when they were young. The subsequent discovery by optical astronomers of extreme 'active galactic nuclei' (quasars) at very large redshifts corroborated Ryle's conjectures, but these objects, and their evolution, are still too poorly understood to be used for determining the geometry of the universe.

By probing deep into space, astronomers can study parts of the universe whose light set out a long time ago. If we lived in a wildly inhomogeneous universe, there would be no reason why these remote regions (and the way they have evolved) should bear any resemblance to our own locality. However, insofar as the universe we find ourself living in (or at least the part of it accessible to observation) is actually uniform and isotropic, its gross kinematics are describable by a single scale factor $R(t)$; all parts have evolved the same way and have the same history (see Figure 2). This simplicity gives us reason to believe that when we observe a region of the universe that lies (say) 3 billion light-years away, its gross features (the statistical properties of the galaxies, the nature of the clustering, etc.) resemble those that would have been displayed 3 billion years ago in our own locality (i.e. within the region depicted in Figure 1).

Astronomers have an advantage over geologists, in that they can directly observe the past. And there has been spectacular progress in the technology for probing faint and distant objects. The first improvement came when photographic plates were replaced by CCD solid-state detectors up to 50 times more sensitive at optical and near infrared wavelengths. The advent of a new generation of telescopes with 10-metre mirror diameters has enhanced astronomers' abilities to study the light from faint objects. (The two Keck Telescopes in Hawaii are already complete; several more are currently being built.)

The faintest and most distant galaxies appear typically only

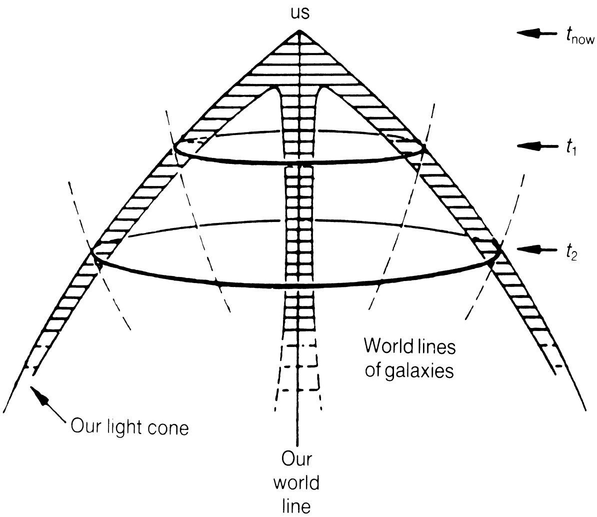

Figure 2

Schematic space-time diagram showing world line of our Galaxy and our past light cone. The only regions of space-time concerning which we have direct evidence are those shaded in the diagram, which lie either close to our own world line (inferences on the chemical and dynamical history of our Galaxy, 'geological' evidence, etc.) or along our past light cone (astronomical evidence). It is *only* because of the overall homogeneity that we can confidently assume any resemblance between the distant galaxies whose light is now reaching us and the early history of our Galaxy. In homogeneous universes we can define a natural time coordinate, such that all parts of the universe are similar on hypersurfaces corresponding to a given value of *t*.

1–3 arcseconds across, and are little more than blurred smudges of light when viewed from the ground, because atmospheric fluctuations smear even a point source over a substantial fraction of an arcsecond. But the Hubble Space Telescope, after its optics was corrected in 1994, has yielded much sharper pictures. The most spectacular single image, the so-called 'Hubble Deep Field', was obtained by pointing the telescope for more than a week towards the same patch of sky.[5a] Observations with this level of sensitivity reveal several hundred galaxies, with a range of morphologies, within a patch only an arcminute square. Redshifts have been measured for many of these, using the Keck Telescope.[5b] In many cases the wavelengths are stretched, between emission and reception, by a factor $R_{now}/R_{em} = 1 + z > 4$: the absorption edge at the Lyman limit (912 Å) is shifted into the visible band, and is indeed the most prominent feature in the spectrum. Larger samples of high-redshift galaxies have been discovered by using this distinctive spectral feature – shifted into the blue part of the visible spectrum – as a diagnostic.[5c]

The light from these remote galaxies set out when the universe was much younger than it is today: we are observing them at a stage when they are only recently formed, and it is not surprising to find that they look distinctively different from nearby systems.

There have been astonishing advances, during the late 1990s, in detecting galaxies at very high redshifts. The observation of high-redshift objects is, however, not in itself so novel: quasars and other 'active galactic nuclei' (e.g. the intense radio sources), the hyperactive centres of a special subset of galaxies, outshine the stellar content of their host galaxy by a factor that can amount to many thousands. These are so bright that high-quality spectra could be taken even with moderate-sized telescopes. An early example of a high-redshift quasar is PC 1247 + 3406, with $z = 4.89$, whose spectrum is shown in Figure 3; the Lyman-α 1216 Å line is observed in the red part of the spectrum, at around 7200 Å. To estimate the relative age of the universe then and now, one needs to know the dynamics of the expansion, and in particular how much it has been decelerating. If there were no deceleration at all, the universe would have been 'younger' when the light set out by the factor $1 + z$ of 5.89.

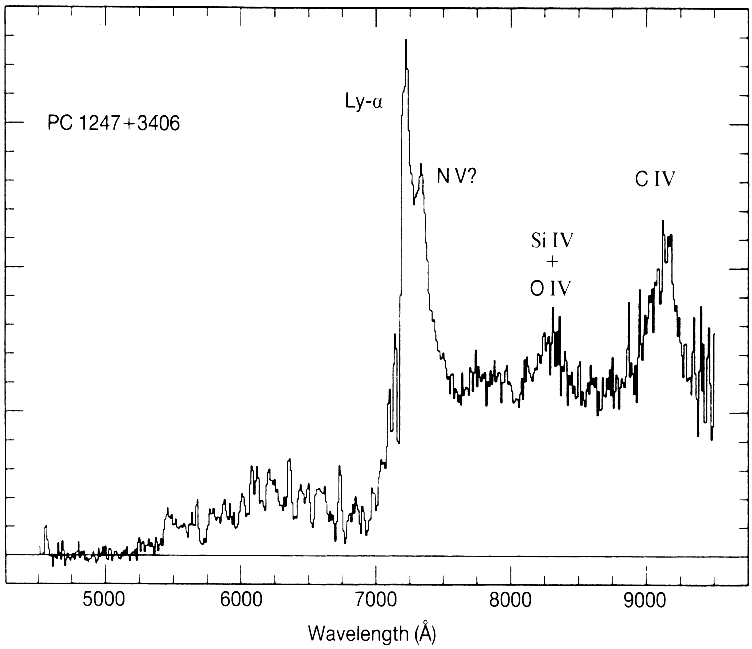

Figure 3

The spectrum of the quasar PC 1247 + 3406, with redshift $z = 4.89$. Light from this object set out towards us when the cosmic scale factor R was $(1 + z) = 5.89$ times smaller than it is today. According to the Einstein–de Sitter model, the universe would then have been only ~ 7 per cent of its present age. (From Schneider, D. P., Schmidt, M. & Gunn, J. E. 1991, *Astron. J.* **102**, 837.)

However, according to the Friedmann models the expansion is decelerating. In the theoretically attractive Einstein–de Sitter cosmology the scale factor of the universe grows as $R \propto t^{2/3}$. The light now reaching us from PC 1247 + 3406, according to that model, would have set out when the universe was younger by a factor $5.89^{3/2}$. Astronomers can therefore probe the last 90 per cent of cosmic history. The existence of these quasars tells us that by the time the universe was about 10^9 years old some galaxies (or at least their inner regions) had already formed, and runaway events in their centres had led to the extreme type of active nuclei that the quasar phenomenon represents.

The host galaxies of quasars should presumably have formed before the quasars themselves; moreover, if galaxies build up hierarchically, smaller galaxies (perhaps themselves too small to host a powerful quasar) should form still earlier. There is therefore every reason to expect galaxies with redshifts substantially larger than 5. These would generally be very faint – certainly too faint for a high-quality spectrum to be obtainable even with a 10 m telescope. However, some faint 'fuzzy' objects with $z > 5$ have been found by extending to higher redshifts the techniques that proved so successful in finding galaxies with $z = 3$.[6a] Another technique for finding them is to use filters for objects whose low-resolution red spectra exhibit a line that is actually highly redshifted Lyman α.[6b] Such attempts have already revealed several galaxies further away than PC 1247 + 3406: in one or two cases the effort is aided by the lucky accident that those galazies are gravitationally lensed (see Chapter 2) by a cluster of galaxies along the line of sight.[6c] It is not clear what the limiting galaxy redshift will be: it depends on how and when galaxy formation starts, a topic discussed further in Chapter 5.

The light from bright quasars offers an important probe for the intervening medium. Absorption lines blueward of Lyman α in the spectrum indicate clouds of gas lying along the line of sight.[6] The absorption is probably caused by protogalaxies too faint to be seen by their direct emission (and where perhaps no stars have yet formed). The way this absorption depends on redshift offers important clues to how galaxy formation proceeded; this is discussed further in Chapter 5.

Pre-galactic history

But what about still earlier epochs, before any galaxies could have formed? Did everything really emerge from a dense (or even singular) 'beginning' ten or fifteen billion years ago? The clinching evidence dates back to 1965, when Penzias and Wilson[7] published their classic paper announcing 'excess antenna temperature at 4080 Mc/s'. Intergalactic space is not completely cold but has a temperature of about 3 K. This may not sound much, but it implies that there are about 4×10^8 photons per cubic metre – maybe a billion photons for every atom in the universe.

The discovery of the microwave background quickly led to a general acceptance of the so-called 'hot-big-bang' cosmology – a shift in the consensus among cosmologists as sudden and drastic as the shift of geophysical opinion in favour of continental drift that took place at about the same time. There seemed no plausible way of accounting for the microwave background radiation except on the hypothesis that it was a relic of an epoch when the entire universe was hot, dense, and opaque. Moreover, the high intrinsic isotropy of the radiation meant that the simple mathematical models were a better approximation to the real universe than the theorists who devised them in the 1920s and 1930s would have dared to hope. Subsequent measurements of this background, made with increasing precision at various wavelengths, have strengthened these conclusions. The radiation spectrum is now known, primarily through the magnificent results from Mather and his collaborators,[8] using the FIRAS (Far Infrared Absolute Spectrophotometer) experiment on the COBE (Cosmic Background Explorer) satellite, to deviate from a black body by less than 1 part in 10^4. The best-fitting temperature is 2.726 K. And measurements by several groups[9–12] show that the radiation is intrinsically isotropic to within a few parts in 10^5, but that there are apparent anisotropies, on angular scales from 0.3° up to 90°, at the 10^{-5} level (some quantitative implications of these are mentioned in Chapter 3).

In the dense early phases, the radiation would have been held in thermal equilibrium with the matter scattering repeatedly off free electrons whose density would have been high enough to make the universe very opaque. But when expansion had cooled the matter below 3000 K (when the cosmic scale factor R was $10^{-3}R_{\text{now}}$) the primordial plasma would have recombined, leaving few free electrons. The 'fog' would then have lifted, the universe thereafter becoming transparent, and probably remaining so until the present (see p. 108). The experimentally detected microwave photons are direct messengers from an era when the universe was about a thousand times more compressed, and had expanded for about half a million years. But the photons are still around – they fill the universe and have nowhere else to go. An important 'cosmic number' is the photon-to-baryon ratio η^{-1}, which stays essentially constant during the cosmic expansion. It is because this ratio is large that many authors refer to the *hot* big bang.

The universe contains other important fossils of a cosmic era far earlier than (re)combination: the light elements such as D, ^{3}He, ^{4}He, and ^{7}Li. During the first *minute* of cosmic expansion, when temperatures were above 10^9 K, nuclear reactions would have synthesised these elements, in calculable proportions, from protons and neutrons. The baryon density in an expanding universe goes as $R^{-3} \propto T^3$, and would therefore have been 10^{27} times higher when $T = 3 \times 10^9$ K than it is today. But this is still not as high as the density of air! One does not need to worry about problems of dense matter. And the energies of the relevant nuclear reactions are <1 MeV, and do not involve any large uncertain extrapolation from the experimental domain. Such calculations,[13] showing how the light-element abundances would depend on the present mean baryon density, the number of neutrino species, etc. were done in the 1960s. Although refinements have been introduced,[14] nothing essential has changed on the theoretical front over the last 25 years.

Stellar nucleogenesis, supernova explosions, and recycling into new stars, the theory of which was formulated in the 1950s,[15] seem well able to account for 'heavy' elements such as carbon, oxygen, and iron.[16] But the high and relatively uniform proportion of helium always posed a problem. It was therefore gratifying, and neatly complementary, that helium was the one element that would be created prolifically in a 'big bang'. In the 1970s, the astrophysical problems of accounting for deuterium (whose abundance is reduced during stellar recycling) were properly appreciated, and this isotope is also believed to be a cosmological fossil.[17]

Only more recently have astronomers been able to determine the light-element abundances in old stars, gaseous nebulae, etc. precisely enough to make a worthwhile comparison with the 'big-bang' predictions. In particular, the helium abundance is now pinned down with a precision approaching 1 per cent. Measurements of deuterium in our own Galaxy yield a lower limit to the primordial abundance, because an uncertain proportion would have been destroyed by processing through earlier generations of stars. It was therefore an important advance when the Keck Telescope allowed astronomers to take such high-quality spectra of quasars that weak lines due to D (displaced from the much stronger H-lines by an isotopic shift equivalent to 80 km s^{-1}) could be measured.[17a] These observations, referring to diffuse gas at early epochs, are likely to give a

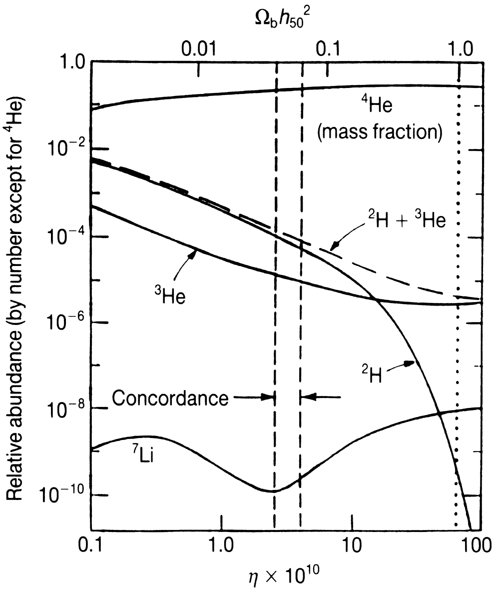

Figure 4

The predicted abundances of the light elements emerging from a standard 'hot big bang', as a function of the baryon/photon ratio η. Note that there is a definite range of η for which the calculations yield abundances of ^{4}He, D, ^{3}He, and ^{7}Li concordant with observations. (From Schramm, D. N. 1991, in *After the First 3 Minutes*, eds. Holt, S. S. *et al.* (American Institute of Physics, New York) p. 12.)

better estimate than local measurements of the actual primodrial abundance of deuterium.

What is remarkable is that, as Figure 4 shows, the light-element abundances all appear concordant with the predictions of 'big-bang nucleosynthesis', provided that the baryon density is in the range 0.1 to 0.3 baryons per cubic metre (a density compatible with what we observe). The measured abundances could have been all over the place, or could have indicated a mean cosmic density that was plainly ruled out; these nucleosynthesis calculations therefore offer a strong vindication for extrapolating a standard big-bang model back to a temperature T such that $kT = 1$ MeV. The grounds for this extrapolation should, I believe, be taken as seriously as, for instance, ideas about the early history of our Earth, which are often based on indirect inferences by geologists and paleontologists that are a good deal less quantitative.

Status of the hot-big-bang hypothesis

I would bet odds of 10 to 1 in favour of the general 'hot-big-bang' concept as a description of how our universe has evolved since it was around 1 second old and at a temperature of 10^{10} K (or ~1 MeV). Some people are even more confident. In a memorable lecture at the International Astronomical Union back in 1982, Zel'dovich[18] claimed that the big bang was 'as certain as that the Earth goes round the Sun'. He must even then have known his compatriot Landau's dictum that cosmologists are 'often in error but never in doubt'!

The case for the standard hot big bang has actually strengthened greatly in the last decade, through better measurements of the background radiation and of the light elements. Moreover, one can think of several discoveries that *could* have refuted the model and which have *not* been made. For instance:

(i) Astronomers might have discovered an object whose helium abundance was zero, or at any rate well below 23 per cent. (Stellar nucleosynthesis can readily enhance helium *above* its pre-galactic abundance, but there seems no astrophysically plausible way of eliminating it.)

(ii) The background radiation spectrum might, as experimental precision improved, have turned out to be embarrassingly different from a black body. In particular, the millimetre-wave background measured by COBE might have been *below* a black-body extrapolation of what had already been reliably determined at centimetre wavelengths. It would not be hard to think of effects that would have added extra radiation at millimetre wavelengths (indeed the smallness of the millimetre excess strongly constrains the input from early star formation, decaying particles, etc.), but it would be hard to interpret a millimetre-wave temperature that was *lower* than a black body fitting the Rayleigh–Jeans part of the spectrum.

(iii) If a stable neutrino had been discovered in the mass range from 100 to 10^6 eV, that would have been incompatible with the standard big-bang model, which would predict about 1.1×10^8 such neutrinos per cubic metre; relic neutrinos would then provide a far higher density in the present universe than is compatible with observations.

These considerations give us confidence in extrapolating right back to the first few seconds of our universe's history and in assuming that the laws of microphysics were the same then as now. Conceivably, this confidence is misplaced, and our satisfaction will prove as transitory as that of a Ptolemaic astronomer who has fitted a new epicycle. But the 'hot big bang' certainly seems vastly more plausible than any equally specific alternative.

If we envisage time on a logarithmic scale, then many important events of cosmic history are being overlooked if we consider only the period $t > 1$ s. Figure 5 depicts how the universe might have evolved right from the Planck time to the present. Uncertainties about the relevant physics impede our confidence in discussing the extensive span of logarithmic time 10^{-43} s $< t < 10^{-4}$ s, when thermal energies exceeded 100 MeV. At

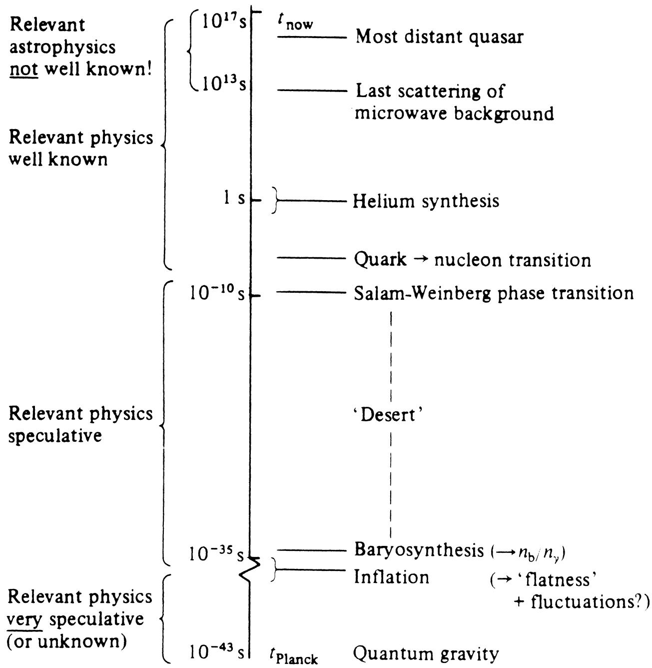

Figure 5

The history of our universe according to the standard hot-big-bang model, showing some of the key physical processes at various stages.

later times, we can consistently use microphysics that is well known; so long as the universe remains almost homogeneous, the evolution is straightforwardly calculable. However, at some stage small initial perturbations must have evolved into gravitationally bound systems (protogalaxies? protoclusters?); the onset of nonlinearity then creates challenging complications, even though the controlling physics is Newtonian gravity and gas dynamics.

Despite the importance and fascination of the ultra-early universe it would be imprudent to venture any bets on what happened when $t \ll 1$ s. The empirical basis for these initial phases of cosmic history is far more tenuous than the quantitative 'fossil evidence' (from light elements and the background radiation) for the eras after 1 second. The first millisecond of cosmic history, a brief but eventful era spanning 40 decades of logarithmic time (starting at the Planck time), is the intellectual habitat of the high-energy theorist and the inflationary or quantum cosmologist. Densities and energies were then so high that the relevant physics is speculative.

From 10^{-3} seconds onwards, quantitative predictions, such as those about cosmic light-element production, are possible; these vindicate our backward extrapolation. (These predictions also, incidentally, vindicate the assumption that the laws of microphysics were indeed the same when the universe had been expanding for only 1 second as they are in our terrestrial laboratories; we should keep our minds open – or at least ajar – to the possibility that this isn't so.)

There has been remarkable progress in the last 25 years in delineating cosmic evolution, mapping out the structure and dynamics of clusters and superclusters, and surveying objects at high redshifts. This progress brings new, strongly interrelated, questions into sharper focus:

(i) How did the dominant present-day structures in our universe – galaxies and clusters – emerge from amorphous beginnings in the early universe?
(ii) What is the dark matter that seems to be the dominant constituent of the universe?
(iii) Are the key parameters that have determined the nature of the present-day universe – the structure, the baryon content, the dark matter, etc. – a legacy of exotic physics in the ultra-early phases?

These lectures are mainly concerned with the first two of these questions (which are interlinked), but the final chapter will touch briefly on current conjectures that relate more directly to the still-mysterious physics governing the earliest phases of cosmic history.